自然语言检索扩展词库的构建方法

吴建荣 陈洪梅 姚建民 熊思勇
（苏州市科学技术情报研究所，江苏苏州 215021）

摘 要：检索词自动扩展词库构建方法的基本思路是：根据语料是否规范化处理进行词库分类建设，优化了系统的检索性能；结合学科类别，对词库语料进行领域划分，引导科技人员对技术领域的准确把握；建设以本体库为基础，将与规范词具有关联性、相似性的语料通过关系表与关联库关联，把科技文献中的关键词组成一个有序的关系网，解决了传统检索系统中检索词无关联的不足；通过对检索词出现频率进行统计分析，进而更新词库，保证本体库、关联库语料的时效性，突破了人工对词库更新管理的受限性。

关键词：自然语言；检索词；检索扩展；本体库；关联库
中图分类号：G354 **文献标识码：**A DOI：10.3772/j.issn.1674-1544.2013.06.013

Lexicon Construction Method for Query Expansion by Natural Language

Wu Jianrong, Chen Hongmei, Yao Jianmin, Xiong Siyong
(Suzhou Institute of Scientific, Technical Information, Suzhou 215002)

Abstract: For high retrieval precision and recall rate, a lexicon construction solution is introduced for query expansion in document retrieval. According to specific technology domain, an ontology based is built on basis of authoritative lexicons by the China national committee for terms and Wiktionary. Synonyms, hypernyms and hyponyms are acquired on basis of template matching and hierarchy structure reasoning from natural language contexts and Wikipedia. For better query expansion performance, a relationship network with statistical link strength is founded on basis of mutual information of related query terms. The above query term network enables a powerful knowledge management tool for document retrieval together with user logs and intermediate retrieval results.

Keywords: natural language, query terms, query expansion, ontology, relation base

1 引言

自然语言是一种自然地随文化演化的语言，是人类交流和思维的主要工具。在信息检索过程中，一般科技人员提交的检索词都具有自然语言特征。相比于自然语言，人工语言是经规范化处理的受控语言，它把表达主题概念的自然语言转换为受控语词进行检索。为了提高检索质量，一般将自然语言转换成人工语言实现准确匹配。本文以苏州市科技服务中心整合同方知网、万方数据、维普资讯等科技文献资源为切入点，研究基于自然语言检索扩展的词库构建方法。

2 词库结构

建设具有逻辑关系扩展的检索词库，是提高检索效率的有效措施。随着资讯、论文、专利等网络资源数量级增长，科研人员为了能准确获取所需的资料，希望被检索的网络资源与自身的研究方向相一致。因此，这里根据科研人员对检索资源的专业领域揭示的要求，并结合同方知网、万方数据、

第一作者简介：吴建荣（1967- ），男，苏州市科学技术情报研究所副所长，副研究员，研究方向：科技管理、科技资源建设与共享、成果转移转化。

基金项目：苏州市2011年基础设施计划项目“苏州市科技文献智能分析公共服务平台”（SZP201107）。
收稿日期：2013年9月28日。

维普资讯等文献资源的主题揭示情况，按学科领域进行关联扩展构建检索词的关联库。涉及的学科类别共34个[1]，如表1所示。同时，以学科类别为基础，利用全国科学技术名词审定委员会公布的名词和全国科学技术名词审定委会员的汉英审定词典规范化的词语作为主题词建立本体库[2]。

在检索过程中，根据用户提交的检索词，以本体库为基础，通过关联词表进行映射，自动抽取出与该词相关或相似的词语，实现检索词扩展。同时，系统对检索词出现频率进行计算，首次出现或在一定阈值以内时存储至关联库，超过一定阈值时以维基词典的关系信息为基础存储至本体库；对于本体库中使用频率低于一定阈值的主题词移至关联库，实现词库的自动更新[3]。如图1所示。

3　本体库

本文选择由全国科学技术名词审定委员会和全国科学技术名词审定委会员汉英审定词典系列公布的词语，以学科类别划分为基础，通过对同义词、近义词、上下位类等关系分析处理形成本体库。同时，以检索词的使用频率为依据，以维基词典的关系信息为基础，自动对本体库进行更新。

3.1　名词

全国科学技术名词审定委员会公布的名词具有权威性和约束力，包括专业术语、术语类别以及术语之间的关系（包括上位词、下位词、同义词等），其中词条数55959条，关系类别数15个，关系实例16365个（即包含相关词的术语个数），实例关系对：57172个，均存储至本体库。例如：“感应分流器”包含如下关系：

<类属>分流器

<子类>多线圈感应分流器

<子类>双线圈感应分流器

<子类>单线圈感应分流器

上述例子中，“关系实例数”为1，“实例关系对”个数为4，每个“关系实例数”包含多个“实例

表1　学科类别信息

A：马列主义、毛泽东思想、邓小平理论	B：哲学、宗教	C：社会科学总论	D：政治、法律
F：经济	G：文化、科学、教育、体育	H：语言、文字	I：文学
J：艺术	K：历史、地理	N：自然科学总论	O：数理科学和化学
Q：生物科学	R：医药、卫生	S：农业科学	T-TB：工业技术
TD:矿业工程	TE:石油、天然气工业	TF:冶金工业	TG:金属学与金属工艺
TH:机械、仪表工业	TJ：武器工业	TK:能源与动力工程	TL：原子能技术
TM:电工技术	TN:无线电电子学、电信技术	TP：自动化技术、计算机技术	TQ：化学工业
TU：建筑科学	TV：水利工程	U：交通运输	V：航空、航天
X：环境科学、安全科学	Z：综合性图书		

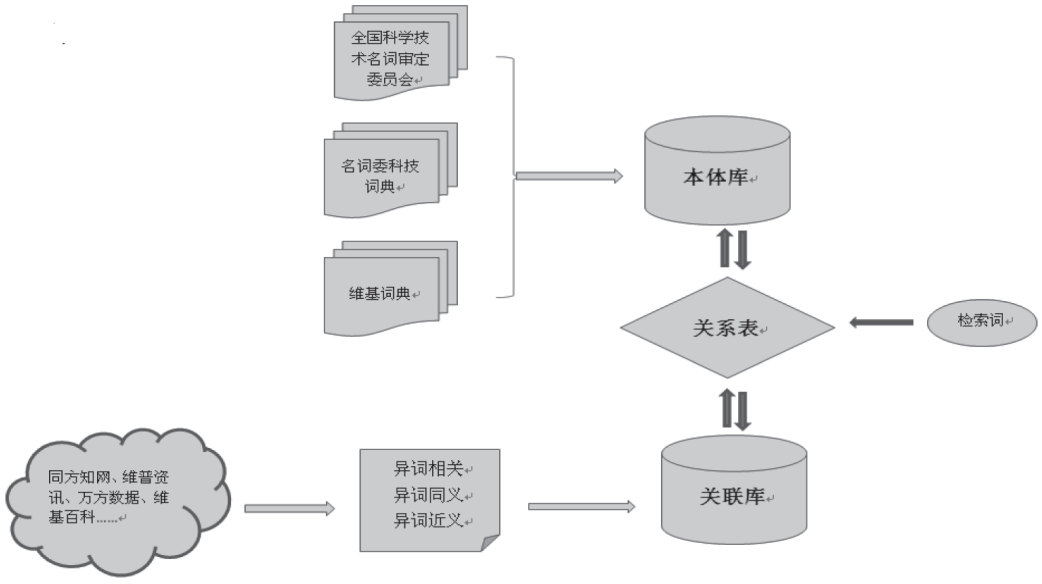


图1　词库结构示意图

关系对”数。

3.2 汉英审定词典

全国科学技术名词审定委会员汉英审定词典包括专业术语以及该术语的英文翻译、上位词、领域和术语定义。图2为术语“作用力”在词典中的组织形式，其中“applied force”是该术语的英文翻译，“机构动力学”为该术语的领域，“机械工程”为该术语的上位词，“能够产生运动或运动趋势的力”是该术语定义。

```
- <Node>
    <id>5927</id>
    <term_ch>作用力</term_ch>
    <term_en>applied force</term_en>
    <step>9</step>
    <systemid>1000</systemid>
    <fieldid>188</fieldid>
    <fieldname>机构动力学</fieldname>
    <innerparentcode>100001840185</innerparentcode>
    <parentname>机械工程</parentname>
    <definitions>能够产生运动或运动趋势的力。</definitions>
    <note />
    <type />
  </Node>
```

图2 汉英审定词典实例

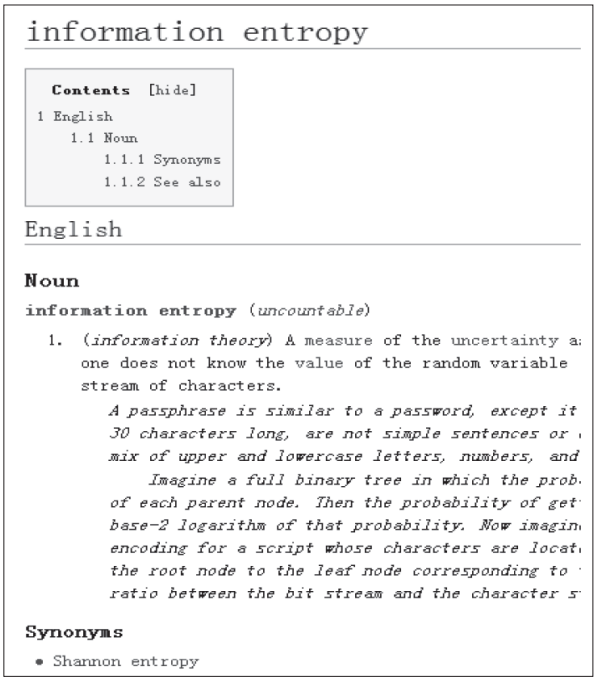
在对全国科学技术名词审定委会员专业术语汉英词典进行抽取时，将其包含的专业术语以及该术语的英文翻译、上位词、领域和术语定义均存储至本体库。

3.3 维基词典

当检索词的检索频率超过一定阈值，本体库未含该检索词信息，且《全国科学技术名词审定委员会》公布名词和《全国科学技术名词审定委员会》的汉英审定词典都没有该检索词的关系信息时，维基词典是一个很好的信息来源。维基词典是一个由志愿者编纂的多语的词典，对一个词汇的发音、语源、释义、词汇翻译给出解释。图3为词条“information entropy（信息熵）”在维基词典中的解释页面。可以看出，“Shannon entropy”为“information entropy”的同义词。维基百科是包含多种语言的词典，其中，英语类词条数最多。目前，通过维基词典共挖掘到包含同义词等相关词语的词条数约32000条。对于新加入本体库的检索词，系统将自动从维基词典中提取关系信息，并经人工筛选确认处理后存储至本体库。

4 关联库

关联库是围绕本体库建立的基于自然语言的词库。关联库中的语料与本体库的语料具有关联关系，是对主题词的扩展，以便科研人员快速地定位


information entropy

Contents [hide]
1 English
1.1 Noun
1.1.1 Synonyms
1.1.2 See also

English

Noun

information entropy (*uncountable*)

1. (*information theory*) A measure of the uncertainty a one does not know the value of the random variable stream of characters.

A passphrase is similar to a password, except it 30 characters long, are not simple sentences or mix of upper and lowercase letters, numbers, and

Imagine a full binary tree in which the prob of each parent node. Then the probability of get base-2 logarithm of that probability. Now imagin encoding for a script whose characters are locat the root node to the leaf node corresponding to ratio between the bit stream and the character s

Synonyms

- Shannon entropy

图3 《维基词典》实例

到相关的研究领域[4]。科研人员针对某个研究领域输入的两个检索词，一般具有异词有关、异词近义、异词同义3种关系[5-7]。

4.1 异词相关

异词相关是指两个不同主题检索词之间具有领域相关性，如，“主题词”与“关键词”。互信息作为一种关联性的度量标准，旨在度量x和y之间的相关程度，其度量公式如下所示[8]：

$$I(x,y)=\log\frac{p(x,y)}{p(x)p(y)} \quad (1)$$

首先根据同方知网、万方数据、维普资讯等文献资源的层次结构，获取这些资源在各领域内论文的题录信息。将各个领域集合中的关键词作为该领域内的关联词集合，再对各领域内的关联词集合分别在标题、关键词和摘要中计算两个词的互信息，其计算公式如下：

$$I(x,y)=\frac{f(x,y)}{f(x)f(y)}\cdot f(x,y) \quad (2)$$

其中，$f(x,y)$为关联词x和关联词y均在标题、摘要或关键词中共现的频度（文章数），$f(x)$为关联词x在标题、摘要或关键词中出现的频度（文章数），$f(y)$为关联词y在标题、摘要或关键词中出现的频度（文章数），该公式在式（1）的基础上再乘以$f(x,y)$是为了防止出现高频词的互信息较低的现象。

将关联词x、y的互信息值，通过标题、关键

词、摘要得到的值分别记为：$I_{标题}(x,y)$、$I_{摘要}(x,y)$和$I_{关键词}(x,y)$，这3个值为关键词相关度度量值。对得到的互信息值采用线性加和的方式将其融合，关联词x以及关联词y的相关度为$\gamma(x,y)$，计算公式如下：

$$\gamma(x,y) = a \cdot I_{标题}(x,y) + b \cdot I_{摘要}(x,y) + c \cdot I_{关键词}(x,y) \quad (3)$$

其中，a、b、c为加权系数，由人工调整。

根据$\gamma(x,y)$相关度排序，将靠前排列的词语作为异词相关进行处理保存。

4.2　异词同义

异词同义是指具有不同描述字符的两个主题检索词表示同一含义，主要体现在同义词、缩写等形式。如“机器翻译”“自动翻译”与“MT”。针对同义相关的检索词，有以下两种方法进行挖掘。

（1）模板匹配法。维基百科对字词具有完整的解释，包括字词的文化背景、文化意义等，这也是维基百科与维基词典的重要区别。该阶段利用维基百科中的词语解释，挖掘具有同一含义的不同词语。例如，对于主题检索词“梯度下降法”，维基百科解释为：“梯度下降法是一个最优化算法，通常也称为最速下降法。”根据该解释，主题检索词“梯度下降法”与“最速下降法”具有同义关系。

（2）词典翻译法。一般来说，一个英文检索词可以被翻译成多个中文词语，如通过有道词典将“information”翻译为中文，可以表示为信息、资料、知识、情报、通知。这里利用有道词典的翻译结果，将具有相同英文翻译的中文检索词判断为同义。

4.3　异词近义

异词近义是指两个不同主题检索词的含义相近，具有上下位关系、包含关系等。如，“概率论”与“概率统计”。针对具有上下位关系的检索词，可通过两种方法实现。

（1）模板匹配法。由于利用自由文本上下位词抽取的准确率低，这里采用模板匹配的方式挖掘上下位关系词。该部分使用的模板通过人工总结，抽取出具有上下位关系的模板定义。根据建立的模板，抽取出现在同一个子句中上下位关系主题词。例如：

<名词“属于”名词“的范畴”>

模板可在论文摘要或其他大规模语料上进行抽取，也可利用关键词两两组合。例如，直接搜索句子“事件抽取属于信息抽取的范畴”，若搜索引擎的返回结果中，能够有完全匹配该句话的结果，或者包含该句话的数量超过某一阈值，则认为“信息抽取”和“事件抽取”具有上下位关系，并且“信息抽取”是“事件抽取”的上位词，“事件抽取”是“信息抽取”的下位词。该部分的抽取旨在补充上一步得到的关键词库中上下位关系信息。

（2）维基百科层次法。维基百科层次法，是指利用维基百科中现有的上下位层次结构，通过同义词扩充，从而得到更多的包含上下位关系的主题检索词。根据图4所示，已知主题检索词A、B为维基百科中的词条，并且两者具有上下位关系，利用同义词构建方法，得到词条A’为词条A的同义词，词条B’为词条B的同义词，那么可以将词条A’以及词条B’加入该上下位关系结构中，从而扩充了上下位关系的主题检索词。

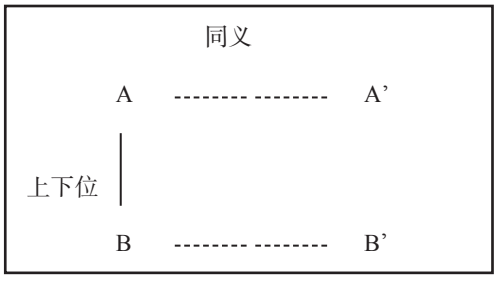


图4　基于维基百科的上下位关系获取

5　结论

本文提出的基于检索词自动扩展的词库构建方法，其基本思路是：根据语料是否规范化处理进行词库分类建设，优化了系统的检索性能；结合学科类别，对词库语料进行主题划分，引导科技人员对技术领域的准备把握；建设以本体库为基础，将与规范词具有关联性、相似性的语料通过关系表与关联库关联，把科技文献中的关键词组成一个有序的关系网，解决了传统检索系统中检索词无关联的不足；通过对检索词出现频率进行统计分析，进而更新词库，保证本体库、关联库语料的时效性，突破了人工对词库更新管理的受限性。

参考文献

[1]　中国图书馆分类法.中图分类号查询[EB/OL].[2013-08-19].http://www.ztflh.com.

[2]　黄媛.基于论文主题词和关键词关系网的检索词扩展

研究[J].科技广场,2011(1):24−27.
[3] 王小华,徐宁,谌志群.基于共词分析的文本主题词聚类与主题发现[J].情报科学,2011,29(11):1621−1624.
[4] 田萱,杜小勇,李海华.信息检索中一种基于词语——主题词相关度的语言模型[J].中文信息学报,2007,21(6):43−50.
[5] 刘华梅.基于情报检索语言互操作技术的集成词库构建研究—以教育词库为例[D].南京:南京农业大学,2006,6.
[6] 王石,曹存根,裴亚军等.一种基于搭配的中文词汇语义相似度计算方法[J].中文信息学报,2013,27(1):7−14.
[7] 梁娜,耿国华,周明全,等.自然语言处理中的语义关系与句法模式互发现[J].计算机应用研究,2008,25(8):2295−2298,2308.
[8] 王夙娟.特定主题词库建立的相关技术的研究[J].科技信息,2012(14):115−116.

国家科技报告服务系统征求意见版正式上线运行

本刊讯 2013年11月1日，国家科技报告服务系统征求意见版正式面向社会上线运行。“国家科技报告服务系统”以推进科技报告资源的开放共享为目的，目前提供在线浏览的1000份科技报告，是依据“十一五”期间已验收的部分国家科技计划项目（课题）验收报告加工而成。科技计划投入所产生的科技报告将通过“国家科技报告服务系统”面向社会开放。公众只要登录网址www.nstrs.cn，就可以了解国家科技计划项目的相关信息。“国家科技报告服务系统”征求意见版的开通，标志着我国科技报告工作全面展开。

科技报告是指科技人员为了描述其从事的科研、设计、工程、试验和鉴定等活动的过程、进展和结果，按照规定的标准格式编写而成的特种文献。科技报告详实记载了项目研究工作的全过程，包括成功的经验和失败的教训，其实质是以积累、传播和交流为目的。科研工作者依据科技报告中的描述能重复实验过程、了解科研结果。科技报告的数量、质量不仅反映了科研项目完成的质量和创新程度，也能验证项目承担人的科研能力和水平，是科研工作承上启下的重要保障。科技报告持续积累所形成的国家基础性战略资源，既为科技管理部门提供真实的信息支撑，又为科研人员提供有效的信息保障，还能保证社会公众对政府科研投入产出的知情权。从而，避免重复投入，实现资源共享。

科技报告试点工作包括4部分内容。一是，要对新老项目实行分类管理。对于已验收的项目，进行科技报告的回溯工作，在提交原有报告基础上，进行科技报告规范改写。对于在研的项目，各计划归口管理部门修改了年度报告、中期报告、验收报告的模板，增加科技报告内容部分。对于新立项目，纳入国家科技计划项目合同管理，计划任务书中将明确规定承担单位呈交科技报告的数量、类型及时限，包括过程中产生的专题技术报告；将科技报告任务完成情况作为中期检查和结题验收的必备条件，作为后续支持的重要依据。二是，在科技部国家科技计划项目申报中心设立科技报告呈交专栏，各科技计划通过相应渠道统一呈交科技报告。同时建设“国家科技报告服务系统”实现公开科技报告的开放共享。三是，积极推进法人单位科技报告体系建设。督促项目（课题）承担单位充分履行法人责任；将科技报告工作纳入本单位科研管理程序，设专门岗位负责科技报告工作，将科技报告纳入机构知识库统一管理；督促项目（课题）负责人组织科研人员撰写科技报告，负责本单位所承担项目（课题）的科技报告审查和呈交工作。四是，由于科研人员不熟悉科技报告格式规范，因此需要对承担国家科技计划课题的科研人员及单位管理人员进行全面培训和宣传工作。以上试点工作正在稳步推进。“国家科技报告服务系统”预计2013年12月底形成总计3000份科技报告的服务规模，2014年3月初完成1万份科技报告上线，面向全社会开放共享。